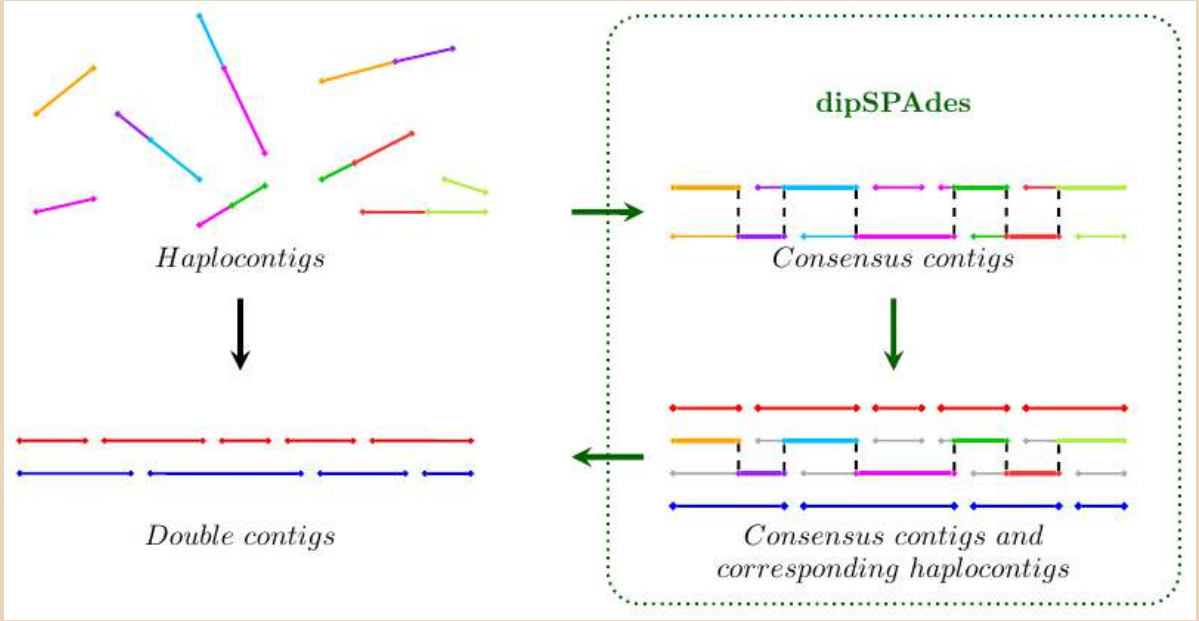
Polymorphism analysis in diploid genomes

Старостина Екатерина
Научные руководители:
Сафонова Яна
Банкевич Антон,
Лаборатория алгоритмической биологии СПБАУ РАН

DipSPAdes

ассемблер для высоко полиморфных диплоидных геномов, использующий граф де Брюйна

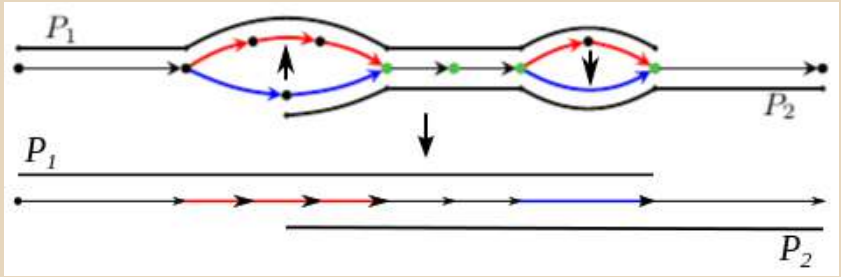


Цель проекта

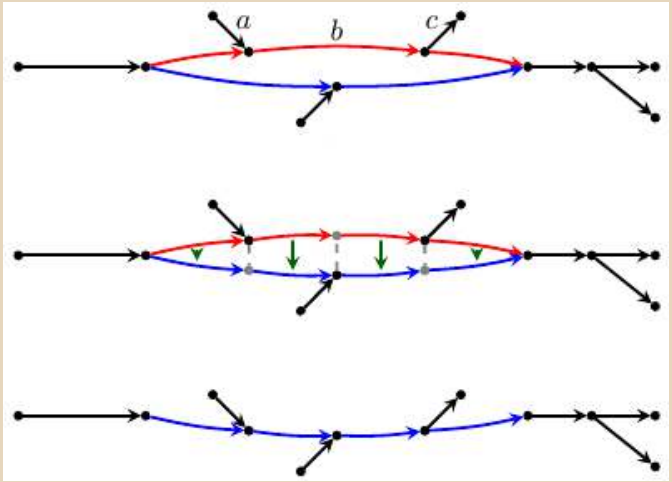
- Анализ сложных полиморфизмов в диплоидном геноме и геномов близкородственных организмов
- Модификация существующего алгоритма построения консенсусных контигов для работы с ними

Основные этапы алгоритма dipSPAdes

1. Построение графа на гаплоконтигах
2. Выравнивание контигов на граф
3. Склеивание пузырей в графе
4. Построение overlap графа на контигах
5. Построение консенсуса



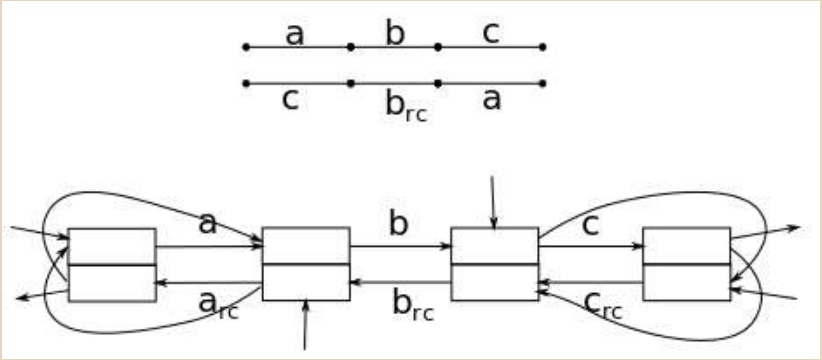
Алгоритм схлопывания полиморфизмов в dipSPAdes



- Для каждой пары контигов ищем образованные ими пузыри в графе
- Выравниваем соответствующие последовательности
- Если последовательности схожи, то схлопываем пузырь

Недостатки алгоритма

- работает только с SNPs и короткими indels
- не склеивает более сложные полиморфизмы, что приводит к разрыву консенсусных контигов в этих местах

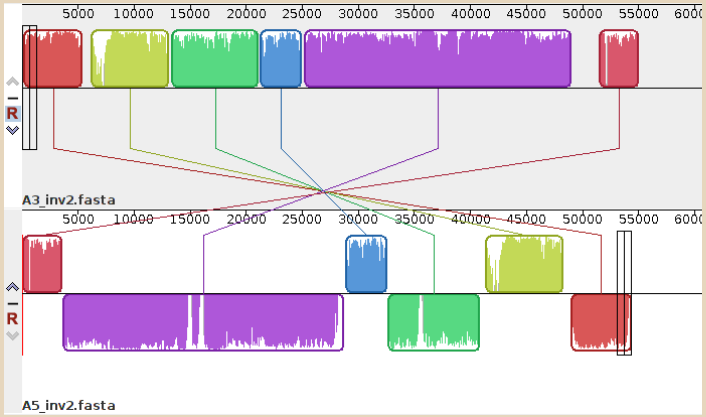


Анализ полиморфизмов

- Симулировали гаплоконтиги как контиги геномов S.commune (уровень полиморфизма ~10%)
- Сделали множественное выравнивание и нашли контиги с инверсиями
- Запустили dipSPAdes на найденных контигах

Проблема

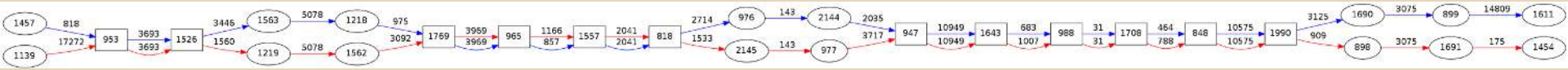
1. Взяты похожие контиги с инверсиями:
2. dipSPAdes строит граф по этим контигам, добавляя комплиментарные
3. Определяются перекрытия контигов по LCS вершин в графе
4. Производится поиск избыточных контигов.



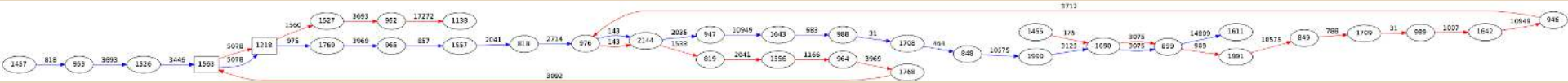
Проблема

Инверсии ⇒

- неправильное выравнивание пузырей
- неправильная длина хвостов путей



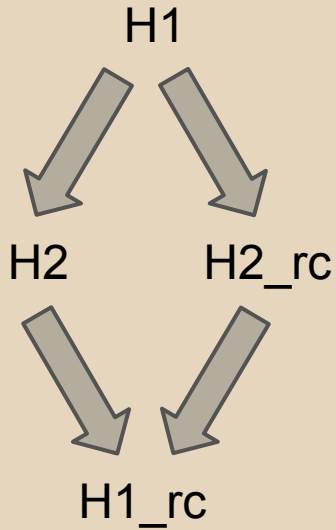
(H1, H2)



(H1, H2_rc)

С точки зрения DipSPAdes (H1, H2) и (H1, H2_rc) перекрываются одинаково

Проблема



Инверсия в контигах приводит к появлению лишних ребер в overlap графе

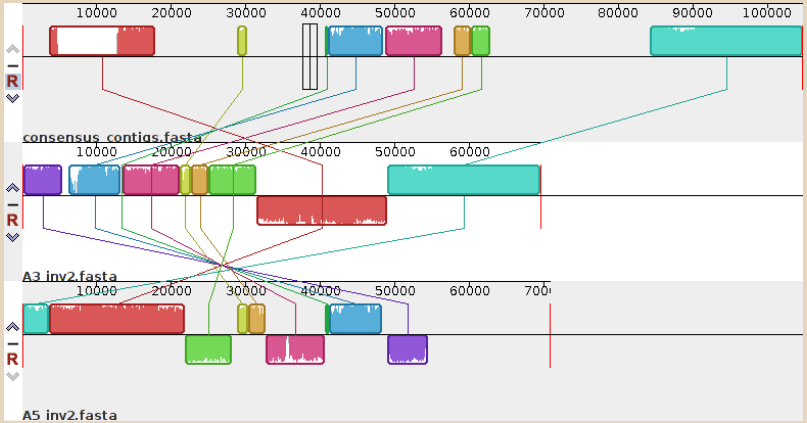
Сделано

- Добавлена корректная обработка инверсий на стадии удаления пузырей и поиска избыточных контигов (схлопывание гомологичных балджей с инверсиями, работа с инверсиями на хвостах путей в графе)
- Добавлена обработка специфических структур в overlap графе, образующихся при наличии инверсий в контигах

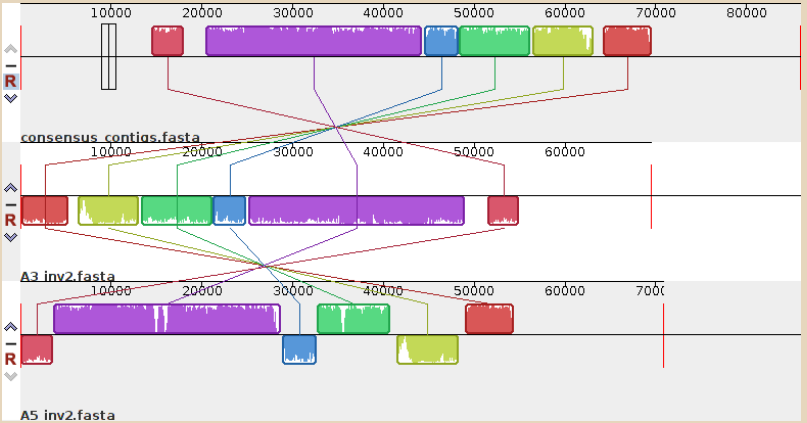
Результаты

- Для пары контигов:

Было:



Стало:



Результаты

- Для сборки консенсуса двух индивидуумов S. commune:

	было	стало
№ контигов	1629	1606
Общая длина	38543779	38388759
Длиннейший контиг	892853	964356
N50	192058	216253
N75	78039	82416

Спасибо за внимание :)